

Clustering Hypertext with Applications to Web Searching

Dharmendra S. Modha and W. Scott Spangler
IBM Almaden, San Jose, CA

IMA Text Mining Workshop
April 17, 2000

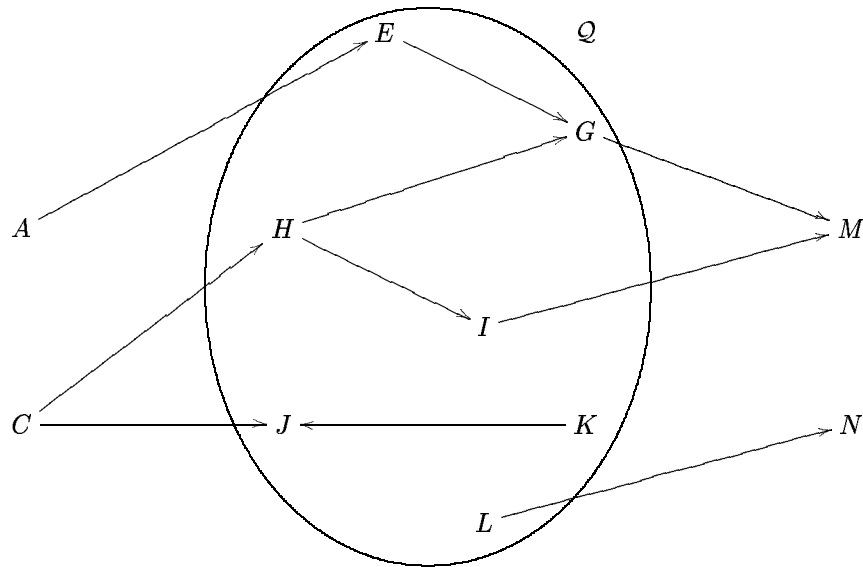
Vector Space Models for Text

(Salton & McGill, 1983)

1. parse documents and construct a **word** dictionary
2. for each document, count the number of occurrences of each **word**
3. discard **stopwords**
4. assign an unique $1 \leq i \leq d$ to remaining **words**
5. for each document, **D** is a **d** -dimensional vector whose i -th component is the number of occurrences of **word** i in the document
6. for each document, normalize **D** to have unit L^2 norm

Vector space models of text are very high-dimensional and very sparse (Dhillon & Modha, 1999).

Vector Space Models for Hypertext



We represent each hypertext document $\mathbf{x} = (\mathbf{D}, \mathbf{F}, \mathbf{B})$, where

- $\mathbf{D} \in R^d$ captures word features
- $\mathbf{F} \in R^f$ captures out-links/forward-links features
- $\mathbf{B} \in R^b$ captures in-links/back-links features

\mathbf{D} is explicit in (Salton & McGill, 1983)

\mathbf{F} and \mathbf{B} are implicit in (Kleinberg, 1997)

Concept Triplet

Given a collection (“cluster”) π of (hyper)text documents, define the *word concept vector* (Dhillon & Modha, 1999) as

$$\mathbf{D}^* = \frac{\sum_{\mathbf{x} \in \pi} \mathbf{D}}{\|\sum_{\mathbf{x} \in \pi} \mathbf{D}\|},$$

where $\mathbf{x} = (\mathbf{D}, \mathbf{F}, \mathbf{B})$ and $\|\cdot\|$ denotes the L^2 norm.

Similarly, define the *out-link and in-link concept vectors* as, respectively,

$$\mathbf{F}^* = \frac{\sum_{\mathbf{x} \in \pi} \mathbf{F}}{\|\sum_{\mathbf{x} \in \pi} \mathbf{F}\|} \text{ and } \mathbf{B}^* = \frac{\sum_{\mathbf{x} \in \pi} \mathbf{B}}{\|\sum_{\mathbf{x} \in \pi} \mathbf{B}\|}.$$

We write *concept triplet* as

$$\mathbf{c}_\pi = (\mathbf{D}^*, \mathbf{F}^*, \mathbf{B}^*)$$

Cluster Annotation

Given a cluster π , and a corresponding concept triplet

$$c_\pi = (\mathbf{D}^*, \mathbf{F}^*, \mathbf{B}^*),$$

we define

- **Summary** as the document triplet $\mathbf{x} = (\mathbf{D}, \mathbf{F}, \mathbf{B})$ that maximizes $\mathbf{D}^T \mathbf{D}^*$.
- **Review** as the document triplet $\mathbf{x} = (\mathbf{D}, \mathbf{F}, \mathbf{B})$ that maximizes $\mathbf{F}^T \mathbf{F}^*$.
- **Breakthrough** as the document triplet $\mathbf{x} = (\mathbf{D}, \mathbf{F}, \mathbf{B})$ that maximizes $\mathbf{B}^T \mathbf{B}^*$.
- **Keywords** as the words corresponding to the largest components of \mathbf{D}^* .
- **Reference** as the out-link corresponding to the largest component of \mathbf{F}^* .
- **Citation** as the in-link corresponding to the largest component of \mathbf{D}^* .

Query: virus

Cluster 0 size = 146

Keywords viruse,anti,software,information,computer,update,antivirus

Summary Anti-Virus Tools (51)

Review SARC Virus EncyclopediaQ - Qm (19)

Breakthrough SARC Virus EncyclopediaXn - Xz (26)

Reference McAfee.com - The Place for Your PC

Citation Zaujimave linky

Query: “human rights”

Cluster 0	size = 157
Keywords	human,international,unit,information,nation,report,law
Summary	Links To Other Human Rights Sources (40)
Review	Derechos Human Rights - contact info (59)
Breakthrough	United Nations Human Rights Website (22)
Reference	Derechos - Human Rights
Citation	Human Rights Reporting: Primary Web Resources

Query: dilbert

Cluster 0 size = 165

Keywords adam,book,scott,comic,work,dogbert,strip

Summary DILBERT ZONE — scott adams past & present (129)

Review DILBERT ZONE — dnrc sock puppets (103)

Breakthrough July 1995: [BUBBA-L:26422] Re: Dilbert (121)

Reference Dilbert Zone

Citation Dilbert : On the Net 700 Sites!

Query: terrorism

Cluster 0 size = 154

Keywords terrorist,state,international,attack,bomb,security,info

Summary US Policy on Terrorism..Part I* (21)

Review Terrorism Research Center: Counterterrorist Org... (34)

Breakthrough Terrorism Research Center: Terrorist Profiles (28)

Reference <http://www.state.gov/www/global/terrorism/>

Citation Terrorism - U.S. News Net Links (116)

A Measure of Similarity

Given two hypertext documents

$$\mathbf{x} = (\mathbf{D}, \mathbf{F}, \mathbf{B})$$

$$\tilde{\mathbf{x}} = (\tilde{\mathbf{D}}, \tilde{\mathbf{F}}, \tilde{\mathbf{B}}),$$

we define

$$S(\mathbf{x}, \tilde{\mathbf{x}}) = \alpha \mathbf{D}^T \tilde{\mathbf{D}} + \beta \mathbf{F}^T \tilde{\mathbf{F}} + \gamma \mathbf{B}^T \tilde{\mathbf{B}}$$

where α , β , and γ are nonnegative numbers such that

$$\alpha + \beta + \gamma = 1.$$

For any triplet $\tilde{\mathbf{x}} = (\tilde{\mathbf{D}}, \tilde{\mathbf{F}}, \tilde{\mathbf{B}})$ and for all (α, β, γ) ,

$$\sum_{\mathbf{x} \in \pi} S(\mathbf{x}, \tilde{\mathbf{x}}) \leq \sum_{\mathbf{x} \in \pi} S(\mathbf{x}, \mathbf{c}_\pi).$$

Toric k -Means Clustering Algorithm

Given n documents

$$\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n,$$

we seek k *disjoint* clusters and corresponding concept triplets

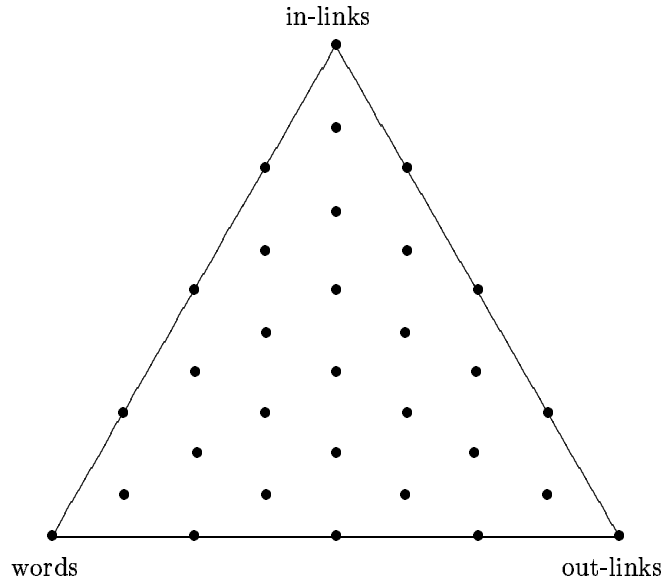
$$(\pi_1, \mathbf{c}_1), (\pi_2, \mathbf{c}_2), \dots, (\pi_k, \mathbf{c}_k)$$

such that the following is maximized:

$$\sum_{j=1}^k \sum_{\mathbf{x} \in \pi_j} S(\mathbf{x}, \mathbf{c}_j).$$

- (1) Select k *initial* concept triplets.
- (2) Recompute clusters by assigning each document to the closest concept triplet (in S).
- (3) Recompute the concept triplets.
- (4) Repeat steps (2) and (3), until convergence.

Choice of the Weights



$$(\alpha^\dagger, \beta^\dagger, \gamma^\dagger) = \arg \max [O_\alpha \times O_\beta \times O_\gamma]$$

where

$$O_\alpha = \left(\frac{\sum_{j=1}^k \sum_{\mathbf{x} \in \pi_j} D^T D_j^*}{\frac{1}{k-1} \sum_{j=1}^k \sum_{\mathbf{x} \in \pi_j} \sum_{\ell=1, \ell \neq j}^k D^T D_\ell^*} \right)^{n_d/n}$$

$\mathbf{x} = (\mathbf{D}, \mathbf{F}, \mathbf{B})$ and n_d is the number of documents with at least one word.

Choice of the Weights...

Query: guinea

α	β	γ	O_α	O_β	O_γ	T
0.990	0.010	0.000	4.20	4.40	3.13	58.18
0.010	0.990	0.000	3.61	6.45	3.24	75.65
0.010	0.000	0.990	3.92	5.92	10.09	234.94
0.010	0.495	0.495	3.73	11.35	7.40	314.55

query	k	α	β	γ
latex	2	0.010	0.000	0.990
abduction	2	0.495	0.010	0.495
guinea	3	0.010	0.495	0.495

Cluster 0	size = 66
Keywords	tex, document, package, command, math, postscript, guide
Summary	Introduction to TeX; LaTeX; BibTeX and SliTeX (78)
Review	TeX and LaTeX (1)
Breakthrough	Peter's TeX/LaTeX/LaTeX2e/LaTeX3 Page (38)
Reference	TeX Frequently Asked Questions
Citation	PROGRAMMING: bookmarks
Cluster 1	size = 82
Keywords	latex, glove, request, allergy, balloon, rubber, product
Summary	Latex Allergy Injuries - The Law Offices Of ... (122)
Review	Enlarger Latex Mattresses - 1(800)FloBeds (188)
Breakthrough	Latex Allergy Injuries - The Law Offices Of ... (122)
Reference	www.FloBeds.com 1(800)FloBeds
Citation	LATEX ALLERGY

Cluster 0 size = 71

Keywords child, children, parent, international, information, court, state

Summary England & Wales - Parental Child Abduction (58)

Review A Halloween Abduction prevention page (105)

Breakthrough Iran - Parental Child Abduction (159)

Reference Islamic Family Law - International Child Abduction (3)

Citation Child Abduction - Divorce Support Net Links

Cluster 1 size = 85

Keywords alien, ufo, story, experience, hip, generator, abductee

Summary Wiendog's Alien Abduction Page (192)

Review What is an alien abduction experience? (116)

Breakthrough Alien Abduction Experience and Research (60)

Reference ABIOGENESIS - POWER OF CREATION

Citation Orthopaedic Rehabilitation. Abduction Pillows. (141)

Cluster 0	size = 92
Keywords	papua,country,png,weather,service,unit
Summary	Papua New Guinea Map (91)
Review	Weather ... Papua ... Forecast (17)
Breakthrough	@datec internet services - Papua ... (46)
Reference	@datec Internet - Papua New Guinea
Citation	Papua New Guinea Orchid News
Cluster 1	size = 34
Keywords	pig,pigs,request,cavy,nance,live,come
Summary	Guinea Pig Links (196)
Review	Todd's Guinea Pig Hutch (6)
Breakthrough	Greg's Guinea Pigs (40)
Reference	Todd's Guinea Pig Hutch (6)
Citation	OinkerNet – Guinea Pigs Worldwide!
Cluster 2	size = 20
Keywords	bissau,travel,information,embassy,island
Summary	Guinea Bissau Travel Notes (70)
Review	Papua New Guinea Travel Notes (160)
Breakthrough	Anthem/Map/People/Economy (23)
Reference	Country Info–The Online Travel Guide
Citation	National Anthems of the World

Modha & Spangler

Paper & Sample Queries

<http://www.almaden.ibm.com/cs/people/dmodha>